УДК 519.7:004.855.5

А. И. Фрей

Комбинаторная оценка вероятности переобучения на основе кластеризации и покрытий множества алгоритмов

Завышенность теоретических оценок обобщающей способности алгоритмов классификации остаётся открытой проблемой уже более сорока лет, начиная с работ В. Н. Вапника и А. Я. Червоненкиса [1]. На практике наиболее перспективным выглядит комбинаторный подход [2], в рамках которого уже удалось добиться улучшения качества логических закономерностей [3]. Данная работа направлена на дальнейшее улучшение качества комбинаторных оценок вероятности переобучения за счет учета сходства между алгоритмами с близкими векторами ошибок.

Рассмотрим задачу классификации. Пусть $\mathbb{X} = \bigl(x_1, \ldots, x_L\bigr)$ — генеральная выборка из L объектов, A — некоторое множество алгоритмов классификации. Пусть $I\colon A \times \mathbb{X} \to \{0, 1\}$ — бинарная функция потерь. Для произвольной подвыборки $U \subseteq \mathbb{X}$ определим *число* и *частоту* ошибок алгоритма $a \in A$, соответственно, как $n(a, U) = \sum\limits_{x_i \in U} I(a, x_i)$ и $\nu(a, U) = n(a, U)/|U|$. *Методом обучения* называют отображение вида $\mu\colon 2^A \times 2^{\mathbb{X}} \to A$. Метод обучения ставит в соответствие множеству алгоритмов A и обучающей выборке $X \subset \mathbb{X}$ некоторый алгоритм $\mu(A, X)$. В данной работе рассматривается метод *пессимистической минимизации эмпирического риска* (ПМЭР), действующий по правилу $\mu(A, X) \in \operatorname*{Argmax}\limits_{a \in A(X)} n(a, \mathbb{X})$ где $A(X) \equiv \operatorname*{Argmin}\limits_{a \in A} n(a, X), \quad \forall X \subset \mathbb{X}$.

Пусть $[\mathbb{X}]^\ell$ — множество всех разбиений генеральной выборки $\mathbb{X}$ на обучающую выборку X длины ℓ и контрольную выборку $\bar{X}$ длины $k = L - \ell$. Для разбиения $\mathbb{X} = X \sqcup \bar{X}$ *переобученностью* алгоритма $a = \mu(A, X)$ называют уклонение частот его ошибок на контроле и на обучении $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$. Следуя [2], определим *вероятность переобучения* $Q_\varepsilon(A, \mathbb{X})$ как долю разбиений $X \sqcup \bar{X}$, при которых переобученность $\delta(\mu(A,\mathbb{X}),X)$ превышает заданный порог $\varepsilon \in (0,1]$:

$$Q_\varepsilon(A,\mathbb{X}) = \mathsf{P}[\delta(\mu(A,X),X) \geqslant \varepsilon], \text{ где } \mathsf{P} \equiv \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell}. \tag{1}$$

Здесь и далее [*истина*] = 1, [*ложь*] = 0.

Введем на A отношение частичного порядка: $a < b$ означает, что $I(a,x) \leqslant I(b,x), \forall x \in \mathbb{X}$ и $a \neq b$. Если $a < b$ и $\exists! x \in \mathbb{X}$ такой, что $a(x) \neq b(x)$, то будем говорить, что a *предшествует* b, и записывать $a \prec b$.

Теорема 1. *Пусть множество алгоритмов A представлено в виде разбиения на непересекающиеся подмножества $A = A_1 \sqcup A_2 \sqcup \dots \sqcup A_t$, такие что внутри каждого A_i алгоритмы допускают равное число ошибок на полной выборке. Пусть μ — ПМЭР. Для каждого A_i рассмотрим порождающее и запрещающее множества X_i и X'_i:*

$$X_i = \bigcap_{a \in A_i} \{x \in \mathbb{X} \colon \exists b \in A \colon a \prec b, I(a,x) < I(b,x)\},$$
$$X'_i = \bigcap_{a \in A_i} \{x \in \mathbb{X} \colon \exists b \in A \colon b < a, I(b,x) < I(a,x)\}.$$

Пусть, кроме этого, каждое подмножество вложено в объемлющее множество: $A_i \subset B_i$, $i = 1, \dots, t$. Тогда

$$Q_\varepsilon(A,\mathbb{X}) \leqslant \sum_{i=1}^{t} P_i \, Q_{\varepsilon_i}(B_i, Y_i), \tag{2}$$

где $P_i = \frac{C_{L_i}^{\ell_i}}{C_L^\ell}$ — верхняя оценка на вероятность $\mathsf{P}[\mu X \in A_i]$, $Y_i = \mathbb{X} \backslash X_i \backslash \bar{X}_i$, $\varepsilon_i = \frac{L_i}{\ell_i k_i} \frac{\ell\, k}{L} \varepsilon + \bigl(1 - \frac{\ell\, L_i}{L \ell_i}\bigr) \frac{m_i}{k_i} - \frac{|X'_i|}{k_i}$, $L_i = L - |X_i| - |X'_i|$, $\ell_i = \ell - |X_i|$, $k_i = k - |X'_i|$, m_i — число ошибок алгоритмов из A_i.

Теорема 1 обобщает метод порождающих и запрещающих множеств [3]. Она позволяет вычислять оценку для семейств с существенно большим числом алгоритмов, т.к. сумма (2) содержит меньшее число слагаемых из-за кластеризации алгоритмов с близкими векторами ошибок. Отметим, что разбиение $A = A_1 \sqcup \dots \sqcup A_t$ и структуру объемлющих множеств B_i можно выбрать произвольно. В частности, можно использовать объемлющие множества с известной точной оценкой вероятности переобучения.

Список литературы

[1] Vapnik V. N., Chervonenkis A. Y. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2), 264–280.

[2] *Воронцов К. В.* Точные оценки вероятности переобучения // Доклады РАН, 2009. — Т. 429, № 1. — С. 15–18.

[3] Vorontsov K. V., Ivahnenko A. A. (2011) Tight combinatorial generalization bounds for threshold conjunction rules. *4-th Int'l Conf. on Pattern Recognition and Machine Intelligence (PReMI'11)*. Lecture Notes in Computer Science, Springer-Verlag, 66–73.